9. HAFTA

DİSKRİMİNANT ANALİZİ

Diskriminant analizi önceden bilinen farklı kitlelerden(gruplardan, sınıflardan, kümelerden) birine, üzerinde ölçüm yapılan yeni bir birimin(bireyin) atanması(sınıflandırılması) biçiminde tanımlanan istatistiksel bir tekniktir.

İki Kitle İçin Fisher Sınıflandırma Yöntemi

Genel düşünce gözlemleri, bilinen iki kitleden veya sınıftan birine dağıtmak ya da yeni bir gözlemi sınıflardan birine atamaktır. Π_1 ve Π_2 sınıfları veya kitleleri göstersin. $\underline{X}' = (X_1, X_2, ..., X_p)$ birey (birim) üzerinden alınacak ölçümlere karşılık gelen p boyutlu rasgele vektör olsun. $\underline{X}$ 'in gözlem değeri kitleden kitleye (gruptan gruba) değişecektir. $\underline{X}$ rasgele vektörüne ilişkin gözlem değeri, $\underline{\mathbf{x}}$, Π_1 ' de ise $\underline{X}$ 'in olasılık yoğunluk fonksiyonu $f_1(\underline{\mathbf{x}})$ ve Π_2 'de ise $\underline{X}$ 'in olasılık yoğunluk fonksiyonu $f_2(\underline{\mathbf{x}})$ dir.

$\underline{\mu}_1 = E(\underline{X} / \Pi_1)$: Π_1 'e ait $\underline{X}$ rasgele vektörünün beklene değeri

$\underline{\mu}_2 = E(\underline{X} / \Pi_2)$: Π_2 'ye ait $\underline{X}$ rasgele vektörünün beklene değeri

ve iki kitle için varyans-kovaryans matrislerinin eşit olduğu kabul edilirse

$$Cov(\underline{X} / \Pi_1) = Cov(\underline{X} / \Pi_2) = Cov(\underline{X})$$
$$\mathbf{\Sigma} = E(\underline{X} - \underline{\mu}_i)(\underline{X} - \underline{\mu}_i)' \ , \ i = 1, 2$$

olmak üzere,

$$Y = \underline{l}'_{px1} \underline{X}_{px1}$$

doğrusal bileşimi göz önüne alınsın. Buradan,

$$\begin{aligned} \mu_{1Y} &= E(Y / \Pi_1) \\ &= E(\underline{l}'\underline{X} / \Pi_1) \\ &= \underline{l}'\underline{\mu}_1 \end{aligned}$$

$$\begin{aligned}\mu_{2Y} &= E(Y/\Pi_2)\\ &= E(\underline{l}'\underline{X}/\Pi_2)\\ &= \underline{l}'\underline{\mu}_2\end{aligned}$$

ve

$$\begin{aligned}\sigma_Y^2 &= Var(Y/\Pi_1,\Pi_2)\\ &= Var(\underline{l}'\underline{X})\\ &= \underline{l}'Cov(\underline{X})\underline{l}\\ &= \underline{l}'\mathbf{\Sigma}\underline{l}\end{aligned}$$

dir. En iyi doğrusal birleşim iki kitle için Y ' nin ortalamaları arası kare uzaklığının, Y' nin varyansına oranlanarak bulunan ifadeyi maksimum yapacak şekilde elde edilmiştir. Yani en iyi doğrusal birleşim

$$\begin{aligned}\frac{(\mu_{1Y}-\mu_{2Y})^2}{\sigma_Y^2} &= \frac{(\underline{l}'\underline{\mu}_1-\underline{l}'\underline{\mu}_2)^2}{\underline{l}'\Sigma\underline{l}}\\ &= \frac{\underline{l}'(\underline{\mu}_1-\underline{\mu}_2)(\underline{\mu}_1-\underline{\mu}_2)'\underline{l}}{\underline{l}'\Sigma\underline{l}}\\ &= \frac{(\underline{l}'\underline{\delta})^2}{\underline{l}'\Sigma\underline{l}}\end{aligned}$$

oranından elde edilir, burada $\underline{\delta}=(\underline{\mu}_1-\underline{\mu}_2)$ iki kitle ortalama vektörlerinin farkıdır. pxp tipindeki $\underline{\delta\delta}'$ matrisi, Π_1 ve Π_2 kitlelerinin ortalamaları arası fark bileşenlerinin kareler ve çapraz çarpımlar toplamı matrisidir.

$\frac{(\underline{l}'\underline{\delta})^2}{\underline{l}'\Sigma\underline{l}}$ ifadesi her $c\neq 0$ için $\underline{l}=c\Sigma^{-1}\underline{\delta}=c\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$ ' nin seçilmesiyle maksimum olur. $c=1$ alınmasıyla elde edilen

$$\begin{aligned}Y &= \underline{l}'_{px1}\underline{X}_{px1}\\ &= (\underline{\mu}_1-\underline{\mu}_1)'\Sigma^{-1}\underline{X}\end{aligned}$$

Lineer birleşimine Fisher'in Lineer Diskriminant Fonksiyonu (LDF) denir. Oranın maksimumu

$$\max_{\underline{l}} = \frac{(\underline{l}'\underline{\delta})^2}{\underline{l}'\Sigma\underline{l}} = \underline{\delta}'\,\Sigma^{-1}\underline{\delta}$$

ile verilir.

Lineer diskriminant fonksiyonu, çok değişkenli Π_1 ve Π_2 kitlelerini öyle tek değişkenli kitlelere dönüştürür ki, bu tek değişkenli kitlelerin ortalamaları arası fark, kitle varyansına göre mümkün olduğunca büyük olsun.

Yeni bir $\underline{x}_0$ gözlemi için diskriminant fonksiyonun değeri $y_0 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x}_0$ biçiminde tanımlansın ve bu lineer birleşime göre oluşan iki tek değişkenli kitlenin ortalamalarının orta noktası

$$\begin{aligned} m &= \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) \\ &= \frac{1}{2}(\underline{l}'\underline{\mu}_1 + \underline{l}'\underline{\mu}_2) \\ &= \frac{1}{2}(\underline{l}'(\underline{\mu}_1 + \underline{\mu}_2)) \\ &= \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2) \end{aligned}$$

dır. Buradan,

$$E(Y_0 / \Pi_1) - m \geq 0$$

ve

$$E(Y_0 / \Pi_2) - m < 0$$

dır. Yani eğer yeni birim $\underline{X}_0$, Π_1'den ise Y_0' ın beklenen değeri orta noktadan büyük, $\underline{X}_0$, Π_2'den ise Y_0' ın beklenen değeri orta noktadan küçük olacaktır. Böylece sınıflandırma kuralı:

$y_0 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x}_0 \geq m$ ise $\underline{x}_0$, Π_1' e atanır

$y_0 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{x}_0 < m$ ise $\underline{x}_0$, Π_2' ye atanır

biçimindedir. Ayrıca aynı sınıflandırma kuralı

$y_0 - m \geq 0$ ise $\underline{x}_0$, Π_1' e atanır

$y_0 - m < 0$ ise $\underline{x}_0$, Π_2' ye atanır

biçiminde de ifade edilebilir.

Burada verilen diskriminant fonksiyonunda kitle parametreleri $\underline{\mu}_1$ $\underline{\mu}_2$ ve Σ kitle parametreleri bilinmektedir. Kitle parametreleri bilinmiyorsa, ilgili kitlelerden alınan örneklemlerden parametreler tahmin edilir. Her bir kitleden alınan n_1 ve n_2 birimlik örneklemlerden sırasıyla kitle parametrelerinin tahmin edicileri $\hat{\underline{\mu}}_1 = \overline{\underline{X}}_1$, $\hat{\underline{\mu}}_2 = \overline{\underline{X}}_2$ ve $\hat{\Sigma} = S_{pooled}$ olmak üzere Fisher'in Örneklem Lineer Diskriminant Fonksiyonu

$$\begin{aligned} Y &= \hat{\underline{l}}' \underline{X} \\ &= (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)' S_{pooled}^{-1} \underline{X} \end{aligned}$$

olarak elde edilir. Burada S_{pooled} iki kitle için birleştirilmiş örneklem varyans kovaryans matrisidir ve $S_{pooled} = S$ alınacaktır.

Böylece iki tek değişkenli kitlenin örneklem ortalama değerleri $\overline{y}_1 = \hat{\underline{l}}' \underline{\mathrm{x}}_1$ ve $\overline{y}_2 = \hat{\underline{l}}' \underline{\mathrm{x}}_2$ arasındaki orta nokta

$$\begin{aligned} \hat{m} &= \frac{1}{2}(\overline{y}_1 + \overline{y}_2) \\ &= \frac{1}{2}(\overline{\underline{\mathrm{x}}}_1 - \overline{\underline{\mathrm{x}}}_2)' \mathrm{S}^{-1} (\overline{\underline{\mathrm{x}}}_1 + \overline{\underline{\mathrm{x}}}_2) \end{aligned}$$

dir ve örneklemler bağlı sınıflandırma kuralı

$y_0 = (\overline{\underline{\mathrm{x}}}_1 - \overline{\underline{\mathrm{x}}}_2)' \mathrm{S}^{-1} \underline{\mathrm{x}}_0 \geq \hat{m}$ ise $\underline{\mathrm{x}}_0$, Π_1' e atanır

$y_0 = (\overline{\underline{\mathrm{x}}}_1 - \overline{\underline{\mathrm{x}}}_2)' \mathrm{S}^{-1} \underline{\mathrm{x}}_0 < \hat{m}$ ise $\underline{\mathrm{x}}_0$, Π_2' e atanır

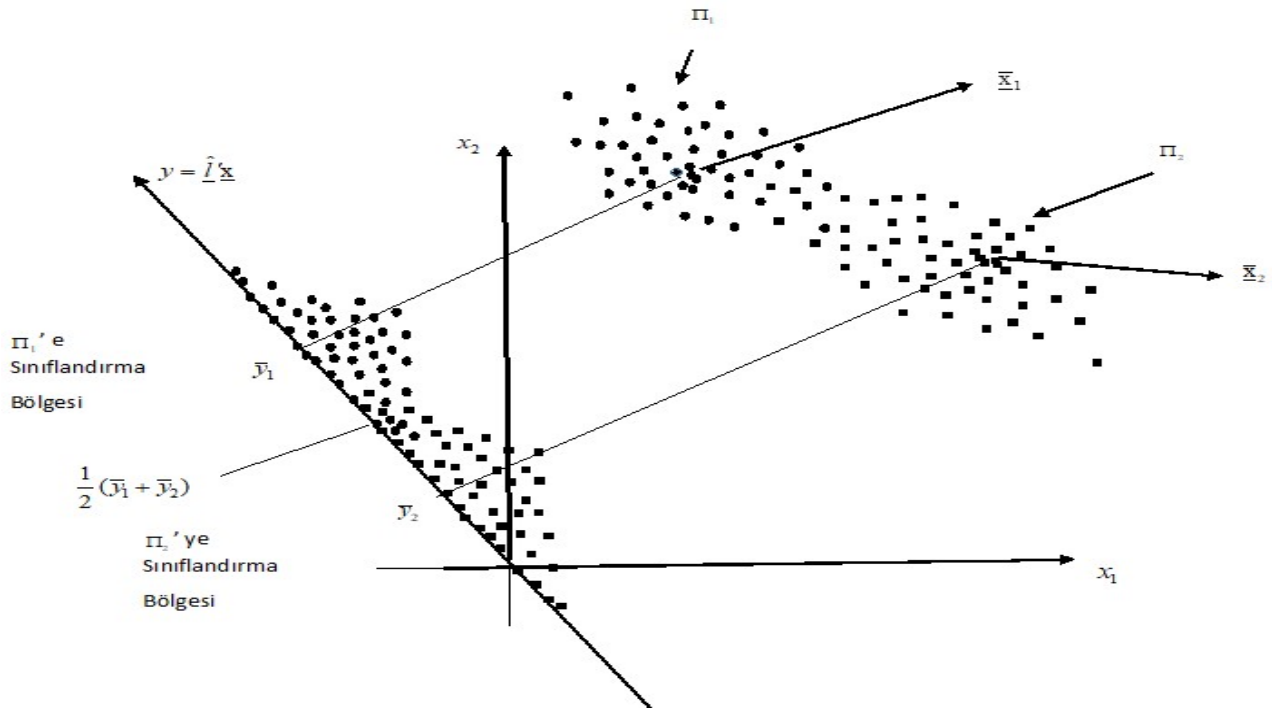
biçiminde ya da

$\mathrm{y}_0 - \hat{m} \geq 0$ ise $\underline{\mathrm{x}}_0$, Π_1' e atanır

$\mathrm{y}_0 - \hat{m} < 0$ ise $\underline{\mathrm{x}}_0$, Π_2' ye atanır

biçiminde de ifade edilebilir.

Ayrıştırma ve sınıflandırma problemi için Fisher çözümü $p = 2$ için aşağıdaki şekilde gösterilmiştir.



Örnek 19 : A tipi Hemophilia hastalığını taşıyan potansiyelin ortaya çıkarılması için kan örnekleri iki kadın grubu için analiz edilmiş ve iki değişken üzerinden

$$X_1 : \log_{10}(AHF\ aktivitesi)$$
$$X_2 : \log_{10}(AHF\ ayni\ antijen)$$

ölçüm değerleri elde edilmiştir. $n_1 = 30$ kadından oluşan ilk grup hastalığı taşımayan grup kitlesinden rasgele seçilmiştir. Bu gruba normal grup adı verilmiştir. $n_2 = 22$ kadından oluşan ikinci grup hemophilia hastası olduğu bilinen kadınların kitlesinden rasgele seçilmiştir.

A tipi taşıyıcılar :hemophilik hastaların kızları, birden çok hemophilik erkek çocuğu olan anneler, bir hemophilik erkek çocuğu ve diğer hemophilik bağıntılı olan anneler. Bu özellikleri taşıyanlara zorunlu taşıyanlar grubu denir.

Ölçümlerden elde edilen örneklem değerleri;

$$\underset{\sim}{\bar{x}}_1 = \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix},\ \underset{\sim}{\bar{x}}_2 = \begin{bmatrix} 0.2483 \\ 0.0262 \end{bmatrix}\ ve\ S^{-1}{}_{pooled} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Buna göre Fisher'in örneklem lineer diskriminant fonksiyonunu, iki tek değişkenli örneklem ortalamalarını ve bunlar arasındaki orta noktayı bulunuz. $x_1 = -0.210,\ x_2 = -0.044$ ölçümlerine sahip bir kadın normal gruba mı, zorunlu taşıyıcılar grubuna mı sınıflandırılır? Her iki normalleştirme metodunu kullanarak orta noktaları elde edip sonucu tekrar yorumlayınız.

Çözüm19:

Π_1: Hastalığı taşımayan grup (bu gruptan alınan örneklem $n_1 = 30$)

Π_2 : Hastalığı taşıyan grup (bu gruptan alınan örneklem $n_2 = 22$)

Fisher'in örneklem lineer diskriminant fonksiyonu :

$$\begin{aligned} y = \underline{\hat{l}}'\underline{x} &= \left[\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right]' S^{-1}{}_{pooled}\,\underline{x} \\ &= \begin{bmatrix} 0.2418 & -0.0652 \end{bmatrix} \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 37.61x_1 - 28.92x_2 \end{aligned}$$

olarak elde edilir. İki tek değişkenli örneklem ortalamaları

$$\bar{y}_1 = \underline{\hat{l}}'\underline{\bar{x}}_1 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.0065 \\ -0.0390 \end{bmatrix} = 0.88$$

$$\bar{y}_2 = \underline{\hat{l}}'\underline{\bar{x}}_2 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.2483 \\ 0.0262 \end{bmatrix} = -10.10$$

Bu ortalamalar arasındaki orta nokta:

$$\hat{m} = \frac{\left(\bar{y}_1 + \bar{y}_2\right)}{2} = -4.61.$$

Fisher'in örneklem lineer diskriminant fonksiyonuna bağlı atama kuralına göre:

Eğer $y_0 = \underline{\hat{l}}'\underline{x}_0 \geq \hat{m} = -4.61$ ise $\underline{x}_0, \Pi_1$'e atanır.

Eğer $y_0 = \underline{\hat{l}}'\underline{x}_0 < \hat{m} = -4.61$ ise $\underline{x}_0$, Π_2'ye atanır.

Soruya dönecek olursak

$$y_0 = \hat{\underline{l}}'\underline{x}_0 = \begin{bmatrix} 37.61 & -28.92 \end{bmatrix} \begin{bmatrix} -0.210 \\ -0.044 \end{bmatrix} = -6.62 < -4.61$$

olduğundan bu kadın zorunlu taşıyıcılar kitlesi olan Π_2 'ye atanır.